Final Report of Traineeship Program 2025

On

"Analysis of Chemical Components"

MEDTOUREASY

27th January 2025

ACKNOWLDEGMENTS

I would like to extend my sincere appreciation to MedTourEasy for providing me with the opportunity to undergo a traineeship in the field of Data Science. This experience has proven to be an invaluable learning journey, allowing me to deepen my understanding of Data Visualizations in Data Analytics and expand my skills both personally and professionally. I am very obliged for having a chance to interact with so many professionals who guided me throughout the traineeship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training & Developement Team of MedTourEasy who gave me an opportunity to carry out my traineeship at their esteemed organization. Also, I express my thanks to the team for making me understand the details of the Data Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for spearing his valuable time in spite of his busy schedule.

I would also like to thank the team of MedTourEasy and my colleagues who made the working environment productive and very conducive.

TABLE OF CONTENTS

ABSTRACT

The beauty and skincare industry faces challenges when it comes to selecting the right cosmetic products, especially for individuals with sensitive skin. Ingredients listed on the back of cosmetic packaging can be difficult to interpret, particularly for those without a background in chemistry. This project aims to simplify the decision-making process for consumers by leveraging Data Science techniques to predict which cosmetic products are best suited for individuals based on their ingredients.

In this project, we leverage Data Science techniques to tackle this problem. The goal is to create a content-based recommendation system that helps consumers make more informed decisions about the cosmetic products they purchase. By analyzing the chemical components of cosmetic products, we use data-driven methods to predict which products are the most likely to suit specific skin types. In particular, the project focuses on the ingredients of moisturizers for dry skin and employs advanced techniques such as word embedding to process these ingredients.

Using a dataset of over 1,472 cosmetics from Sephora, the project performs data preprocessing, including tokenizing ingredient lists and creating a document-term matrix (DTM). We then apply t-SNE (t-Distributed Stochastic Neighbor Embedding) to reduce the dimensionality of the data, allowing us to visualize ingredient similarities between products. The final output is a highly interactive Bokeh visualization that allows users to explore the similarity of products and make informed decisions based on their skin type and ingredient preferences.

This project not only demonstrates the power of machine learning and data visualization in providing actionable insights to consumers but also emphasizes how Data Science can be applied to real-world challenges to improve user experiences in industries like cosmetics.

I. INTRODUCTION

1.1 About the Company

MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. MedTourEasy provides analytical solutions to our partner healthcare providers globally.

1.2 About the Project

Buying new cosmetic products can be a challenging and often overwhelming process, especially for individuals with sensitive skin. The information required to make an informed decision is usually found on the ingredient lists, but these lists can be difficult to understand without a background in chemistry. This project aims to simplify the decision-making process by leveraging Data Science techniques to predict which cosmetic products are most suitable for consumers based on their ingredients.

With a growing number of consumers concerned about skincare, it is crucial to identify products that cater specifically to their needs. However, for individuals with sensitive skin, selecting the right product can often feel like trial and error. Instead of relying on uncertain methods, Data Science provides an opportunity to remove the guesswork and assist consumers in making better choices.

This project focuses on creating a content-based recommendation system that uses the chemical ingredients of each cosmetic product as the key factor in making predictions. The project involves processing the ingredient lists of 1,472 cosmetic products from Sephora using word embedding, followed by visualizing ingredient similarities using t-SNE.

Interactive visualizations will be created using Bokeh, which will allow consumers to explore the relationships between products based on their ingredients.

The goal of this project is to provide consumers with more accurate and personalized product recommendations, particularly for those with sensitive or dry skin. By analyzing product ingredients, this project aims to reduce the trial and error process and guide users toward the most suitable products for their skin type.

The project is divided into the following key subsections:

- *Analysis of the data:* This step involves importing and inspecting the data to understand its structure. We will focus on identifying product categories and skin types, and filter the data to focus on moisturizers for dry skin.

- *Ingredient Tokenization:* Ingredients will be tokenized to create a document-term matrix (DTM), which forms the foundation of the recommendation system.

- *Data Preprocessing:* The data will be processed by applying techniques like one-hot encoding to prepare it for analysis.

- *Dimension Reduction with t-SNE:* Using t-SNE, the dimensions of the ingredient data will be reduced to two, enabling easier visualization.

- *Visualization and Interaction:* A Bokeh plot will be used to map the products based on ingredient similarities, providing an intuitive and interactive dashboard for users.

- *Comparison of Products:* To help consumers further, the project will allow comparisons between two similar cosmetic products to examine ingredient differences.

By the end of this project, users will be able to gain insights into which products are best suited for their skin, providing them with an intuitive and data-driven method for selecting cosmetics.

1.3 Objectives and Deliverables

This project is focused on building interactive, dynamic, and easily interpretable dashboards using data from various cosmetic sources. By leveraging techniques from data science, including natural language processing (NLP), the project aims to analyze and recommend cosmetics based on their ingredients, helping consumers make informed decisions. The project will utilize coding languages such as Python, with libraries like Pandas, NumPy, Scikit-learn, and Bokeh for creating visualizations.

The main objectives and deliverables of the project are outlined below:

a. Ingredient-based Recommendation System: This part of the project aims to analyze and provide product recommendations based on ingredient similarities. It includes:

- Analyzing the ingredients of cosmetic products to determine their compatibility with different skin types.
- Tokenizing the ingredients and building a matrix of ingredient occurrences across different products.
- Creating a recommendation system that predicts which products are likely to suit a user's skin based on the similarity of their ingredient list.
- Using machine learning techniques such as t-SNE for dimensionality reduction to visualize ingredient similarities in a 2D space.
- Providing visual representations using Bokeh to interactively map cosmetic products based on ingredient similarity.

b. Product Similarity Visualization: The project will include a comprehensive analysis of the similarities between various cosmetic products in terms of ingredients. This dashboard will feature:

- Interactive visualizations showing the clustering of products with similar ingredients.
- A tool that allows users to compare two products based on their ingredient lists, helping users find products with comparable formulations.
- A graphical representation of ingredient clusters, visualized through t-SNE, making it easy to understand the relationship between various cosmetic items.

c. User Interaction and Product Exploration:
This dashboard aims to allow users to explore cosmetic products in a more detailed and personalized manner, which includes:

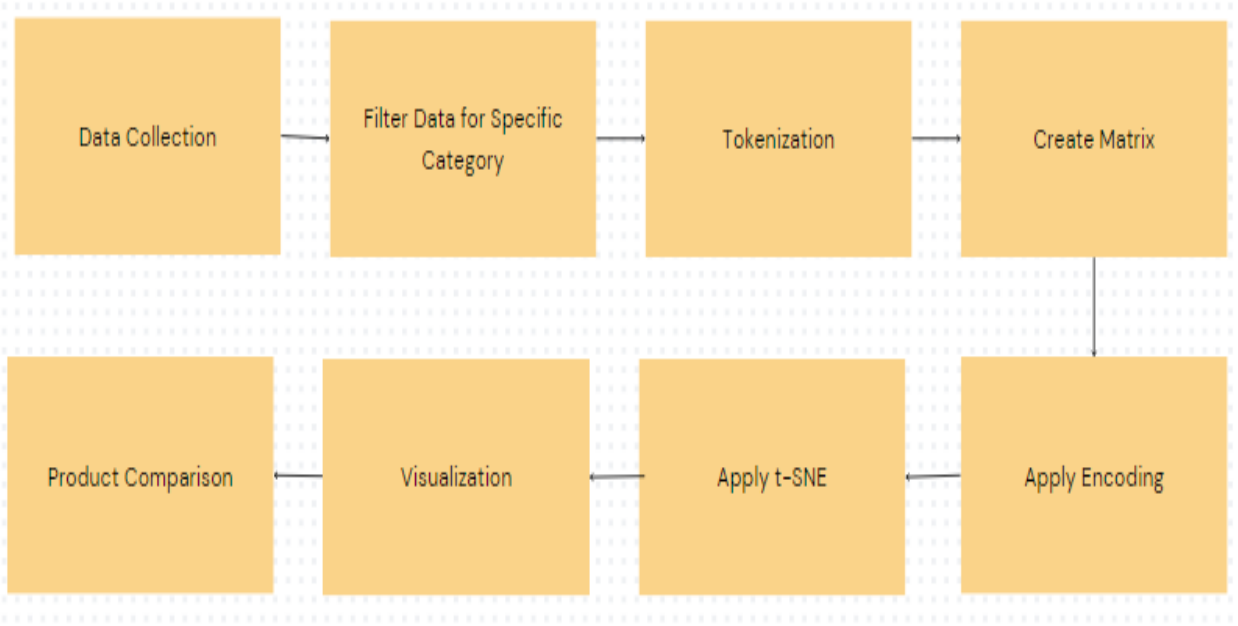
- Mapping of cosmetic products based on user preferences and ingredient compatibility.
- Displaying product details such as brand, price, rank, and ingredients in an interactive dashboard.
- Enabling users to compare and explore different products based on various parameters such as brand and price.
- Adding hover tools and interactivity to the visualization for a more engaging experience.

Each of these objectives will be delivered as interactive and informative dashboards, which can be used by consumers and cosmetics companies to make more data-driven decisions regarding product selection and marketing.

II. METHODOLOGY

2.1 Flow of the Project

The project followed the following steps to accomplish the desired objectives and deliverables. Each step has been explained in detail in the following section.



2.2 Language and Platform Used

2.2.1 Language: Python

For this project, Python is the primary programming language. Python's extensive support for data science, machine learning, and data visualization makes it the ideal choice for this project. We leverage Python to process the ingredient data of cosmetic products, create recommendation models, and visualize the results interactively.

Key Features of Python:

- *Machine Learning Libraries:* Python's scikit-learn library is used for performing machine learning tasks, including tokenizing ingredient lists, creating a document-term matrix (DTM), and applying the t-SNE technique for dimensionality reduction.
- *Data Processing:* Python provides pandas for efficient data manipulation and preparation, making it easier to work with the cosmetic ingredient dataset.
- *Visualization:* Python's Bokeh and matplotlib are used to create visualizations, including interactive scatter plots to map cosmetic items based on ingredient similarity.

2.2.2 IDE: Jupyter Notebook

Jupyter Notebook serves as the development environment for this project. It allows for an interactive workflow, making it easy to analyze and visualize data in steps. This environment is particularly useful for testing, running code snippets, visualizing plots inline, and documenting the analysis process.

Key Features of Jupyter Notebook:

- *Interactive Data Analysis:* Enables step-by-step execution of code and immediate visualization of results.
- *Inline Plotting:* Supports embedding visualizations directly within the notebook, which is essential for presenting t-SNE plots and ingredient similarities.
- *Documentation:* Allows for combining code and explanatory text in markdown, facilitating clear communication of the project's progress

2.2.3 Package: pandas, scikit-learn, Bokeh, numpy

Several Python packages are employed for different aspects of the project.

- pandas is used to load, filter, and manipulate the cosmetic ingredient data.
- scikit-learn provides the machine learning tools for tokenizing ingredients, creating the document-term matrix (DTM), and applying t-SNE for dimensionality reduction.

- Bokeh is used for creating interactive visualizations, including scatter plots, and implementing hover tools for user interaction.
- numpy is used for numerical operations, helping in matrix manipulations like the creation of one-hot encodings and performing mathematical calculations required for machine learning tasks.

Key Features of These Packages:

- *pandas*: Data manipulation (filtering, grouping, and reshaping), ideal for handling CSV data.
- *scikit-learn:* Tokenization, creating the document-term matrix, dimensionality reduction with t-SNE, and general machine learning tasks.
- *Bokeh:* Interactive visualizations with hover functionality for product comparison.
- *numpy*: Efficient array handling and numerical operations essential for matrix creation and calculations.

2.2.4 Template: Bokeh + t-SNE for Interactive Visualization

Bokeh and t-SNE are combined to visualize the similarity between cosmetic products based on their ingredients. t-SNE reduces the high-dimensional ingredient data to 2D space, and Bokeh allows for an interactive plot where users can visually inspect clusters of similar products. The hover tool in Bokeh provides additional details, such as product name, brand, price, and rank, when users hover over any data point.

2.2.5 Dynamic element: t-SNE + Bokeh Hover Tool

The t-SNE algorithm is used to reduce the high-dimensional ingredient data of each cosmetic product into two dimensions. The reduced data is then visualized using Bokeh, with hover tools that show additional product details, such as the product name, brand, price, and rank.

III. IMPLEMENTATION

3.1 Gathering Requirements and Defining Problem Statement

Buying new cosmetic products can be a daunting task, particularly for individuals with sensitive skin who need to carefully consider ingredient lists before making a purchase. The challenge lies in interpreting the ingredient lists that often contain scientific names and terms, making it difficult for the average consumer to assess which products are suitable for their skin type.

In this project, the objective is to develop a content-based recommendation system to aid consumers in selecting cosmetics based on their ingredients. Specifically, we focus on predicting which products may be a good fit for individuals with dry skin by analyzing the ingredients of products categorized as moisturizers. The recommendation system will leverage a machine learning technique, using word embeddings to process ingredient lists, reduce dimensionality via t-SNE (t-distributed Stochastic Neighbor Embedding), and visualize the ingredient similarities using interactive plots created with Bokeh.

3.2 Data Collection and Importing

Dataset Overview

The dataset used for this project is sourced from a CSV file that contains details of cosmetic products from Sephora. The dataset consists of 1,472 rows and the following columns:

- **Label:** Product category (e.g., Moisturizer)
- **Brand:** Brand name of the product
- **Name:** Name of the product
- **Price:** Price of the product
- **Rank:** Product ranking
- **Ingredients:** A list of ingredients used in the product
- **Skin Types:** Product suitability for various skin types (Combination, Dry, Normal, Oily, Sensitive

Data Importing and Inspection:

The dataset was imported using pandas into a DataFrame named df. The first few rows were displayed for inspection using df.sample(), and the counts of product types were checked with df['Label'].value_counts() to understand the distribution of different products.

3.3 Data Cleaning

While the dataset provided contains most of the necessary information, some cleaning steps are required before proceeding with analysis:

1) Handling Missing Data:

No missing data was found in the key columns ,so no imputation was needed.

2) *Filtering Relevant Data:*

Only moisturizers suitable for dry skin were filtered from the dataset. This was done by selecting products where the Label column contains "Moisturizer" and filtering further by dry skin suitability, indicated by the Dry column

3) *Ingredient Tokenization:*

Ingredients are tokenized by converting the ingredient list into lowercase text and splitting by commas. Special handling is implemented to avoid redundancy in ingredient names.

3.4 Data Representation (Document-Term Matrix)

After tokenizing the ingredients, a Document-Term Matrix (DTM) is created to represent the frequency of each ingredient across all products in the filtered dataset. Each row represents a product, and each column represents a specific ingredient.

The total number of products is assigned to M and the total number of unique ingredients is assigned to N. A matrix of size M x N is initialized to store the one-hot encoded values of ingredients for each product.

3.5 Dimensionality Reduction with t-SNE

To visualize the ingredient similarity, we used t-SNE to reduce the dimensions of the DTM from high-dimensional space to two dimensions. The resulting 2D feature matrix is then added as new columns to the moisturizers_dry DataFrame for visualization.

3.6 Data Visualization

The reduced data was visualized using Bokeh, creating an interactive scatter plot. The X and Y axes represent the two dimensions obtained from the t-SNE model. Products are plotted as circles with different tooltips, displaying additional information like the product name, brand, price, and rank.

For this content-based recommendation system, evaluating the results is primarily visual. The scatter plot allows us to observe clusters of similar products. Products with similar ingredients are located closer together in the plot, indicating that the system has effectively captured ingredient similarity.

IV. SAMPLE SCREENSHOTS AND OBSERVATIONS

This section presents the sample screenshots from the interactive visualizations generated by the content-based recommendation system and outlines the observations made during the process.

4.1 Dataset Inspection

The initial dataset, cosmetics.csv, was imported and inspected to understand the structure and variety of products available. The dataset consists of various product categories, ingredient lists, and suitability for different skin types.

Key Observations:

- The dataset includes 1,472 cosmetics categorized under multiple labels such as moisturizers, serums, and cleansers.
- The Label column reveals the count of each product type using value_counts(). Moisturizers constitute a significant proportion of the dataset, making them an ideal focus for this project.

	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive
1393	Sun protect	SUPERGOOP!	Super Power Sunscreen Mousse Broad Spectrum SP...	34	4.3	Aluminum Starch Octenylsuccinate, Ananas Sativ...	1	1	1	1	0
1241	Eye cream	DERMADOCTOR	Wrinkle Revenge Eye Balm	50	4.1	Water, Caprylic/Capric Triglyceride, Palmitoyl...	0	0	0	0	0
1345	Sun protect	SUPERGOOP!	Defense Refresh Setting Mist Broad Spectrum SP...	28	3.3	-Avobenzone 3%, Homosalate 4%, Octisalate 3%, ...	1	1	1	1	0
812	Treatment	KATE SOMERVILLE	EradiKate™ Mask Foam-Activated Acne Treatment	54	4.2	Water, Sodium C14-16 Olefin Sulfonate, Coco-Gl...	0	0	0	0	0
674	Treatment	CAUDALIE	Resveratrol Lift Firming Serum	82	4.3	Water, Glycerin*, Butylene Glycol, Glyceryl St...	1	1	1	1	1

```
Label
Moisturizer    298
Cleanser       281
Face Mask      266
Treatment      248
Eye cream      209
Sun protect    170
Name: count, dtype: int64
```

4.2 Filtering for Moisturizers and Dry Skin

To align the analysis with a specific product category and skin type:

- The dataset was filtered for products labeled as "Moisturizer".
- Further filtering was performed to retain only moisturizers suitable for dry skin.

Key Observations:

- The filtered dataset, moisturizers_dry, focuses on moisturizers specifically suited for dry skin, reducing the dataset size for targeted analysis.
- Resetting the index ensures the dataset is prepared for further processing.

4.3 Tokenizing Ingredients

To compare products based on their ingredient lists:

- Each product's ingredient list was tokenized into individual components.
- A dictionary, ingredient_idx, was created to map unique ingredients to numerical indices for encoding.

4.4 Document-Term Matrix and One-Hot Encoding

A document-term matrix (DTM) was initialized, where:

- Rows correspond to products.
- Columns represent the presence (1) or absence (0) of specific ingredients.

One-hot encoding was applied to populate the matrix using the custom oh_encoder function.

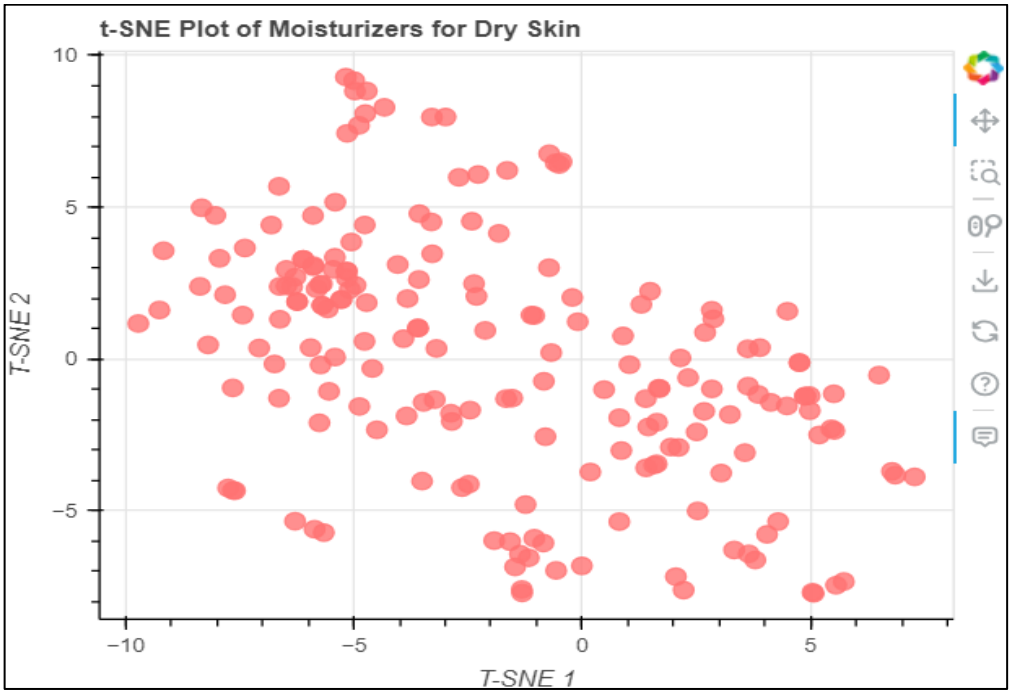
Key Observations:

- The DTM provides a binary representation of ingredient presence for each product.
- It is the foundation for measuring ingredient-based similarity between products.

4.5 Dimensionality Reduction with t-SNE

Using t-SNE, the high-dimensional DTM was reduced to two dimensions for visualization:

- The fit_transform method applied to the matrix produced two features: X and Y.
- These features were added as columns to the filtered dataset, moisturizers_dry.



4.6 Visualization with Bokeh

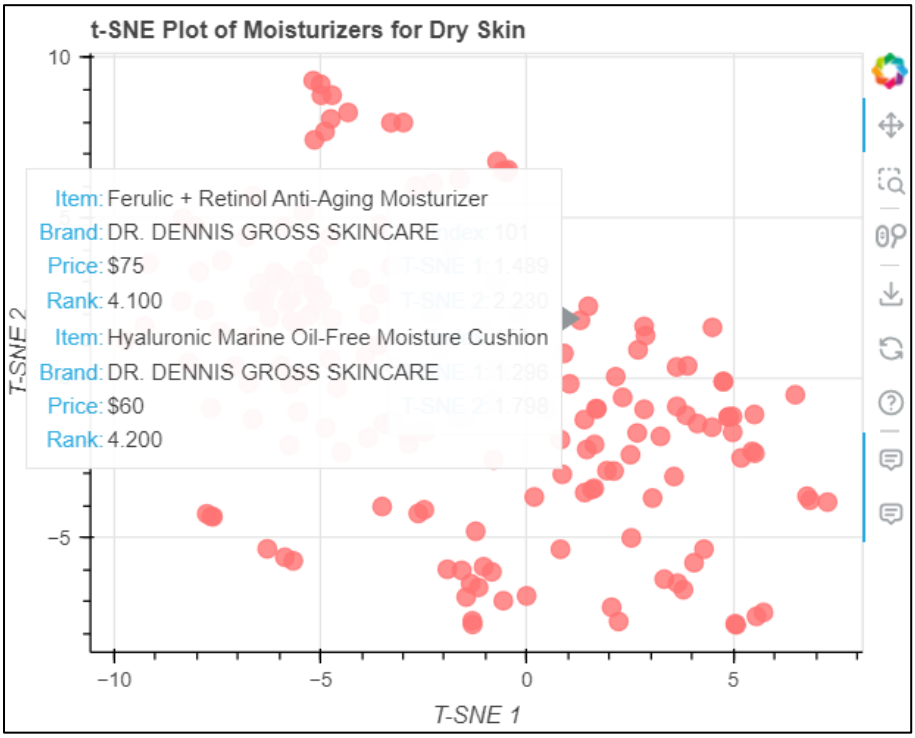
An interactive scatter plot was created using the Bokeh library:

- Each point represents a moisturizer for dry skin.
- Hover functionality was added to display product details, including Name, Brand, Price, and Rank.

Key Observations:

- Similar products are grouped into clusters based on ingredient similarity.

- Hovering over a point provides additional context, enabling quick comparisons between products.



4.7 Comparing Similar Products

Two closely clustered products, AMOREPACIFIC Color Control Cushion Compact Broad Spectrum SPF 50+ and LANEIGE BB Cushion Hydra Radiance SPF 50, were compared based on their ingredient lists:

Key Observations:

- Both products share common active ingredients such as Titanium Dioxide, making them effective sunscreens.
- Differences in inactive ingredients likely account for variations in price and user ratings.

	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive	X	Y
45	Moisturizer	AMOREPACIFIC	Color Control Cushion Compact Broad Spectrum S...	60	4.0	Phyllostachis Bambusoides Juice, Cyclopentasil...	1	1	1	1	1	-1.320351	-7.601377

['Phyllostachis Bambusoides Juice, Cyclopentasiloxane, Cyclohexasiloxane, Peg-10 Dimethicone, Phenyl Trimethicone, Butylene Glycol, Butylene Glycol Dicaprylate/Dicaprate, Alcohol, Arbu

	Label	Brand	Name	Price	Rank	Ingredients	Combination	Dry	Normal	Oily	Sensitive	X	Y
55	Moisturizer	LANEIGE	BB Cushion Hydra Radiance SPF 50	38	4.3	Water, Cyclopentasiloxane, Zinc Oxide (CI 7794...	1	1	1	1	1	-1.313422	-7.716459

['Water, Cyclopentasiloxane, Zinc Oxide (CI 77947), Ethylhexyl Methoxycinnamate, PEG-10 Dimethicone, Cyclohexasiloxane, Phenyl Trimethicone, Iron Oxides (CI 77492), Butylene Glycol Dic

V. CONCLUSION AND FUTURE SCOPE

The content-based recommendation system developed in this project demonstrated the power of data science in simplifying the selection of cosmetic products tailored to specific skincare needs. By focusing on ingredient similarities, the system effectively clustered products and provided actionable insights, especially for users with sensitive skin or targeted requirements. The use of t-SNE for dimensionality reduction enabled clear and intuitive visualizations, while Bokeh's interactive capabilities enhanced the user experience by offering detailed product information, such as name, brand, price, and rank. The analysis of moisturizers for dry skin showcased the system's ability to identify closely related products and compare them based on their chemical compositions. This framework highlights how ingredient-based recommendations can empower consumers to make more informed decisions, reducing the uncertainty involved in purchasing skincare products.

Looking ahead, there are several avenues to enhance and expand this recommendation system. First, the approach can be extended to other product categories, such as serums, cleansers, and sunscreens, while adapting the system to address the needs of different skin types, including oily and combination skin. Integrating user reviews and ratings could complement the ingredient-based analysis by providing valuable insights into product performance and customer satisfaction. Further, advanced natural language processing techniques, such as word embeddings, could deepen the understanding of ingredient relationships and improve the accuracy of recommendations. Including factors like price, availability, and brand reputation would increase the system's practical relevance for users. Finally, deploying the system as a mobile app or web-based tool would make personalized recommendations widely accessible, transforming the way consumers approach cosmetic product selection.

VI. REFERENCES

Data Collection

The following websites have been referred to obtain data for the cosmetic recommendation:

- Sephora – Cosmetics Data
- FDA - Cosmetic Ingredient Information
- Cosmetics Info – Ingredient Information
- Kaggle Dataset: Cosmetics Products
- Beautypedia – Skincare Reviews

Programming References

The following websites have been referred to for programming tutorials, libraries, and tools related to data science, machine learning, and visualization:

- Scikit-learn Documentation
- Pandas Documentation
- TensorFlow Tutorials
- Bokeh Documentation
- t-SNE – A Machine Learning Dimensionality Reduction Method
- Jupyter Notebook
- Kaggle – Data Science and Machine Learning Tutorials